CLAIMS

What is claimed is:

1. A method for determining whether records are similar in a database containing both structured and unstructured, free-text data, the method comprising the steps of:

accessing two of the records from the database for evaluation; and

evaluating a match between the two records as a weighted match between each of a plurality of available fields, such that a matching process is selected as appropriate from among a group of matching processes including strict Boolean, ordinal, and vector-based matching processes, wherein:

when a strict Boolean matching process is selected, applying a match function as an exact match test;

when an ordinal matching process is selected, applying a match function that makes use of information concerning the size and ordering of the data domain; and

when a vector-based matching process is selected applying a match function that uses a vector space frequency test.

2. The method of claim 1 wherein the step of evaluating a match between the two records comprises applying the matching process to determine a match score for two corresponding fields of the plurality of available fields, the two corresponding fields selected from corresponding locations in each of the two records.

3. The method of claim 1 wherein the step of evaluating a match between the two records comprises selecting the matching process based on a common data type shared by both of two fields of the plurality of available fields accessed in the two records.

4. The method of claim 3 wherein when a Boolean matching process is selected, the data type of both of the two fields specifies nominal data.

5. The method of claim 3 wherein when an ordinal matching process is selected, the data type of both of the two fields specifies data capable of being ordered.

6. The method of claim 3 wherein, when a vector-based matching process is selected, the data type of both of the two fields specifies text data.

7. The method of claim 1 wherein the step of evaluating the match between the two records comprises calculating a similarity score between the two records, as follows:

$$sim(record_i, record_j) = w_1*match(a_{1i},a_{1j}) + w_2*match(a_{2i},a_{2j}) + \ldots w_n*match(a_{ni},a_{nj})$$

wherein sim is a similarity function that determines the similarity score for the two records;

$record_i$ is a first record of the two records and is identified in the database by an iterator i;

$record_j$ is a second record of the two records and is identified in the database by an iterator j;

iterator n identifies a field position for a given field a_{ni} in the $record_i$ and a corresponding field position for a given field a_{nj} in the $record_j$;

match indicates the match function; and

a symbol w_n indicates a predefined weight for each result of each match function.

8. The method of claim 1 wherein the database is a relational database, the records are tuples, and the fields are attributes.

9. A data processing system for determining whether records are similar in a database containing both structured and unstructured, free-text data, the data processing system comprising:

a communications interface for communicating with the database; and

a processor coupled to the communications interface, the processor hosting and executing a data evaluation application that is configured to:

access two of the records from the database for evaluation; and

evaluate a match between the two records as a weighted match between each of a plurality of available fields, such that a matching process is selected as appropriate from among a group of matching processes including strict Boolean, ordinal, and vector-based matching processes, wherein:

when a strict Boolean matching process is selected, apply a match function as an exact match test;

when an ordinal matching process is selected, apply a match function that makes use of information concerning the size and ordering of the data domain; and

when a vector-based matching process is selected, apply a match function that uses a vector space frequency test.

10. The data processing system of claim 9 wherein the data evaluation application is configured to apply the matching process to determine a match score for two corresponding fields of the plurality of available fields, the two corresponding fields selected from corresponding locations in each of the two records.

11. The data processing system of claim 9 wherein the data evaluation application is configured to select the matching process based on a common data type shared by both of two fields of the plurality of available fields accessed in the two records.

12. The data processing system of claim 11 wherein when the data evaluation application selects a Boolean matching process, the data type of both of the two fields specifies nominal data.

13. The data processing system of claim 11 wherein when the data evaluation application selects an ordinal matching process, the data type of both of the two fields specifies data capable of being ordered.

14. The data processing system of claim 11 wherein, when the data evaluation application selects a vector-based matching process, the data type of both of the two fields specifies text data.

15. The data processing system of claim 9 wherein the data evaluation application is configured to calculate a similarity score between the two records, as follows:

$sim(record_i, record_j) = w_1*match(a_{1i},a_{1j}) + w_2*match(a_{2i},a_{2j}) + \ldots w_n*match(a_{ni},a_{nj})$

wherein sim is a similarity function that determines the similarity score for the two records;

$record_i$ is a first record of the two records and is identified in the database by an iterator *i*;

$record_j$ is a second record of the two records and is identified in the database by an iterator *j*;

iterator *n* identifies a field position for a given field a_{ni} in the $record_i$ and a corresponding field position for a given field a_{nj} in the $record_j$;

match indicates the match function; and

a symbol w_n indicates a predefined weight for each result of each match function.

16. The data processing system of claim 9 wherein the database is a relational database, the records are tuples, and the fields are attributes.